

On the Consistency of AUC Optimization

Wei Gao, Zhi-Hua Zhou*

*National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China*

Abstract

AUC (area under ROC curve) is an important evaluation criterion, which has been popularly used in diverse learning tasks such as class-imbalance learning, cost-sensitive learning, learning to rank and information retrieval. Many learning approaches are developed to optimize AUC, whereas owing to its non-convexity and discontinuousness, almost all approaches work with surrogate loss functions. Therefore, the study on AUC consistency is crucial, and the previous study showed that classification calibration is necessary and sufficient for the consistency of AUC.

In this paper, we show that, for pairwise surrogate loss of AUC, minimizing the expected risk over the whole distribution is not equivalent to minimizing the conditional risk on each pair of instances. We disclose that classification calibration is necessary yet insufficient for AUC consistency, and provide a new sufficient condition for the asymptotic consistency of learning approaches based on surrogate loss functions. Based on this finding, we prove that exponential loss, logistic loss and distance-weighted loss are consistent with AUC. Then, we derive the *q-norm hinge loss* and *general hinge loss* that are consistent with AUC. We also derive the consistent bounds for exponential loss and logistic loss, and obtain the consistent bounds for many surrogate loss functions under the non-noise setting. Furthermore, we disclose an equivalence between the exponential surrogate loss of AUC and exponential surrogate loss of accuracy, and one straightforward consequence of such finding is that AdaBoost and RankBoost are equivalent. *Key words:* AUC, consistency, surrogate loss, cost-sensitive learning, learning to rank

*Corresponding author. Email: zhouzh@nju.edu.cn

1. Introduction

AUC (area under ROC curve) is an important evaluation criterion which exhibits strong robustness to the change of class distribution, and thus can be adopted even when classical criteria such as *accuracy*, *precision*, *recall*, etc. are inadequate [PFK98, PF01]. It has been widely used in many learning tasks such as cost-sensitive learning, class-imbalance learning, learning to rank, information retrieval, etc. [Elk01, FISS03, CM04, BBB⁺07, AM08, CV09, CVD09, RS09, KDH11, FHOR11].

Owing to its non-convexity and discontinuousness, it is not easy, or even infeasible, to optimize AUC directly since such direct optimization often leads to NP-hard problem. Instead, surrogate loss functions are usually optimized, such as exponential loss [FISS03, RS09] and hinge loss [BS05, Joa05, ZHJY11]. Minimizing such losses is generally easy, and can be done in polynomial time. An important question then is how well does minimizing such convex surrogate losses lead to minimizing the actually AUC; in other words, does the expected risk of learning with surrogate loss functions converge to the Bayes risk of AUC? Consistency (also called Bayes consistency) guarantees that optimizing a surrogate loss will yield ultimately an optimal function with Bayes risk. Thus, the above problem, in a formal expression, is whether the optimization of surrogate loss functions is consistent with AUC.

1.1. Our Contribution

Previous study shows that classification calibration is necessary and sufficient for the consistency of AUC, whereas we find that it ignores an important prerequisite, that is, for pairwise surrogate loss of AUC, minimizing the expected risk over the whole distribution is not equivalent to minimizing the conditional risk on each pair of instances. We prove that classification calibration is necessary yet insufficient for AUC consistency, e.g., hinge loss and absolute loss are classification-calibrated whereas they are inconsistent with AUC. We further provide a new sufficient condition for the asymptotic consistency of learning approaches based on surrogate loss functions. Based on this finding, we prove that exponential loss, logistic loss and distance-weighted loss are consistent with AUC. Then, we derive the *q-norm hinge loss* and *general hinge loss* that are consistent with AUC. We also derive the consistent bounds for exponential loss and logistic loss, and obtain the

consistent bounds for many surrogate loss functions under the non-noise setting. Further, we disclose an equivalence between the exponential surrogate loss of AUC and exponential surrogate loss of accuracy, and one straightforward consequence of such finding is that AdaBoost and RankBoost are equivalent.

1.2. Related Work

The studies on AUC can be traced back to 1970's in signal detection theory [Ega75], and it has been widely used as a criterion in medical area and machine learning [PFK98, PF01, Elk01], especially for model selection where AUC exhibits a better measure than accuracy theoretically and empirically [HL05]. AUC can be estimated under parametric [ZOM02], semi-parametric [HT96] and non-parametric [HM82] assumptions, and the non-parameteric estimation of AUC is popularly applied in machine learning and data mining, equivalent to the Wilcoxon-Mann-Whitney (WMW) statistic test of ranks [HM82]. Hand [Han09] and Flach et al. [FHOR11] gave the incoherent and coherent explanations of AUC as a measure of aggregated classifier performance, respectively.

AUC has also been regarded as an performance criterion for information retrieval and learning to rank, especially for bipartite ranking [CSS99, FISS03, CM04, RS09, Rud09]. Various Generalization bounds are presented to understand the prediction beyond the training sample [AGH⁺05, UAG05, CMR07, CLV08, AN09, RS09]. Also, the learnability of AUC has been studied by Agarwal and Roth [AR05]. More recently, Kotlowski et al. [KDH11] introduced univariate surrogate loss functions to optimize bipartite ranking.

Breiman [Bre04] initiated the consistency issue and showed that exponential loss converges to the Bayes classifier for arcing-style greedy boosting algorithms in the infinite sample case. Buhlmann and Yu [BY03] studied the consistency of boosting algorithms with respect to least square loss. The consistent theory for support vector machines are developed in [Lin02, Ste05], and the influential and fundamental work for binary classification has been investigated comprehensively by Zhang [Zha04b] and Bartlett et al. [BJM06], in which many famous algorithms (e.g., boosting, logistic regression and SVMs) are proved to be consistent. Furthermore, the consistent theory on multi-class classification has been addressed in [Zha04a, TB07] and many SVM-style algorithms are proved to be inconsistent. More recently, the consistency on multi-label learning has been

studied by Gao and Zhou [GZ11]. Much attention has also been paid to the consistency analysis on learning to rank [CZ08, XLW⁺08, XLL09, DMJ10].

In contrast to previous studies on consistency [Zha04a, Zha04b, BJM06, TB07, GZ11] that focused on single instances, our work concerns about the surrogate loss functions of AUC that focused on a pair of instances from different classes. This crucial difference leads to the fact that in contrast to previous studies that are sufficient to focus on conditional risk, our study on AUC consistency has to consider the whole distribution, because as to be shown in Lemma 1, minimizing the expected risk over the whole distribution is not equivalent to minimizing the conditional risk. This is a challenge for the study on AUC consistency. Previous study [CLV08, Section 7 pp. 864] suggested to analyze the AUC consistency by directly extending the results of Bartlett et al. [BJM06], i.e., classification calibration is necessary and sufficient for AUC consistency. Our study shows that classification calibration is necessary yet insufficient for AUC consistency. Kotlowski et al. [KDH11] and Agarwal [Aga12] studied AUC via minimization univariate losses, which is different from the pairwise surrogate losses of our concern.

Duchi et al. [DMJ10] studied the consistency of supervised ranking, but it is quite different from our work. Firstly, the problem settings are different: they considered “instances” consisting of a query, a set of inputs and a weighted graph, and the goal is to order the inputs according to the weighted graph; yet we consider instances with positive or negative labels, and the goal is to rank positive instances higher than negative ones. Moreover, they established inconsistency for the logistic loss, exponential loss and hinge loss even in low-noisy setting, yet our work shows that the logistic loss and exponential loss are consistent but hinge loss is inconsistent.

Rudin and Schapire [RS09] established the equivalence between AdaBoost and RankBoost in the asymptotic behavior (iteration number converges to infinity) for finite training sample when the negative and positive classes contributed equally. In Section 5, we derive an equivalence between the exponential surrogate loss of AUC and the exponential surrogate loss of accuracy when the size of training sample approaches to infinity; this provides a new explanation to the asymptotic equivalence between AdaBoost and RankBoost.

1.3. Organization

Section 2 introduces some preliminaries and previous studies on AUC consistency. Section 3 shows that classification calibration is necessary yet insufficient for AUC consistency, and we present a new sufficient condition. Section 4 studies consistent bounds. Section 5 discloses the equivalence between the exponential surrogate losses of AUC and accuracy. Section 6 presents detailed proofs. Finally, Section 7 concludes and raise some open problems.

2. Preliminaries

Let \mathcal{X} denote an instance space and $\mathcal{Y} = \{+1, -1\}$ the label set. We denote by \mathcal{D} an unknown (underlying) distribution over $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{D}_{\mathcal{X}}$ represents the instance-marginal distribution over \mathcal{X} . For convenience, the conditional probability $\eta: \mathcal{X} \rightarrow [0, 1]$ is defined as

$$\eta(\mathbf{x}) = \Pr[y = +1 | \mathbf{x}].$$

We consider a training sample of n_1 positive instances and n_2 negative instances

$$S = \{(\mathbf{x}_1, +1), \dots, (\mathbf{x}_{n_1}, +1), (\mathbf{x}'_1, -1), \dots, (\mathbf{x}'_{n_2}, -1)\}$$

drawn identically and independently according to distribution \mathcal{D} . Let $f: \mathcal{X} \rightarrow \mathbb{R}$ be a score function. Then, the AUC with respect to sample S and function f is defined as

$$\text{AUC}(f, S) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(I[f(\mathbf{x}_i) > f(\mathbf{x}'_j)] + \frac{1}{2} I[f(\mathbf{x}_i) = f(\mathbf{x}'_j)] \right),$$

where $I[\cdot]$ is the indicator function which returns 1 if the argument is true and 0 otherwise.

Optimizing the AUC is equivalent to minimizing the empirical risk

$$\hat{R}(f, S) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \ell(f, \mathbf{x}_i, \mathbf{x}'_j),$$

where the loss function $\ell(f, \mathbf{x}_i, \mathbf{x}'_j) = I[f(\mathbf{x}_i) < f(\mathbf{x}'_j)] + [f(\mathbf{x}_i) = f(\mathbf{x}'_j)]/2$ is also called *ranking loss*. It is easy to get

$$\text{AUC}(f, S) + \hat{R}(f, S) = 1.$$

We define the expected risk of function f as $R(f) = E_S[\hat{R}(f, S)]$, which is equivalent to

$$R(f) = E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_{\mathcal{X}}^2} [\eta(\mathbf{x})(1 - \eta(\mathbf{x}'))\ell(f, \mathbf{x}, \mathbf{x}') + \eta(\mathbf{x}')(1 - \eta(\mathbf{x}))\ell(f, \mathbf{x}', \mathbf{x})]. \quad (1)$$

Denote by the Bayes risk

$$R^* = \inf_f [R(f)],$$

where the infimum takes over all measurable functions. By simple calculation, we can get the set of optimal functions, also called *set of Bayes predictors*:

$$\mathcal{B} = \{f: R(f) = R^*\} = \{f: (f(\mathbf{x}) - f(\mathbf{x}'))(\eta(\mathbf{x}) - \eta(\mathbf{x}')) > 0 \text{ if } \eta(\mathbf{x}) \neq \eta(\mathbf{x}')\}. \quad (2)$$

Notice that the ranking loss ℓ is non-convex and discontinuous, and a direct optimization often leads to NP-hard problems. In practice, surrogate loss functions that can be optimized with efficient algorithms are usually adopted. Throughout this paper, we consider the following formulations of pair-wise surrogate loss functions:

$$\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}')),$$

where ϕ is a convex function, e.g., exponential loss $\phi(t) = e^{-t}$ [FISS03, RS09], hinge loss $\phi(t) = \max(0, 1 - t)$ [BS05, Joa05, ZHJY11], etc. Similarly, we define the expected ϕ -risk as

$$R_\phi(f) = E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_{\mathcal{X}}^2} [\eta(\mathbf{x})(1 - \eta(\mathbf{x}'))\phi(f(\mathbf{x}) - f(\mathbf{x}')) + \eta(\mathbf{x}')(1 - \eta(\mathbf{x}))\phi(f(\mathbf{x}') - f(\mathbf{x}))], \quad (3)$$

and denote by the optimal expected ϕ -risk,

$$R_\phi^* = \inf_f R_\phi(f)$$

where the infimum takes over all measurable functions.

Many notions on consistency have been introduced in the literature, e.g., the Fisher consistency [Lin02], infinite-sample consistency [Zha04a], classification calibration [BJM06, TB07], edge-consistency [DMJ10], multi-label consistency [GZ11], etc. In this paper, we define formally the *AUC consistency* as follows:

Definition 1 *The surrogate loss ϕ is said to be consistent with AUC if for every sequence $\{f^{(n)}(\mathbf{x})\}_{n \geq 1}$, the following holds over all distributions \mathcal{D} on $\mathcal{X} \times \mathcal{Y}$:*

$$R_\phi(f^{(n)}) \rightarrow R_\phi^* \text{ then } R(f^{(n)}) \rightarrow R^*.$$

For two given instances \mathbf{x} and \mathbf{x}' , we define the conditional ϕ -risk as

$$C(\mathbf{x}, \mathbf{x}', \alpha) = \eta(\mathbf{x})(1 - \eta(\mathbf{x}'))\phi(\alpha) + \eta(\mathbf{x}')(1 - \eta(\mathbf{x}))\phi(-\alpha), \quad (4)$$

where $\alpha = f(\mathbf{x}) - f(\mathbf{x}')$, and we have

$$R_\phi(f) = E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_X^2} [C(\mathbf{x}, \mathbf{x}', \alpha)].$$

For convenience, we denote by $\eta = \eta(\mathbf{x})$ and $\eta' = \eta(\mathbf{x}')$. Then, we define the optimal conditional ϕ -risk

$$\begin{aligned} H(\eta, \eta') &= \inf_{\alpha \in \mathbb{R}} C(\mathbf{x}, \mathbf{x}', \alpha) \\ &= \inf_{\alpha \in \mathbb{R}} \{ \eta(1 - \eta')\phi(\alpha) + \eta'(1 - \eta)\phi(-\alpha) \}, \end{aligned}$$

and further define

$$H^-(\eta, \eta') = \inf_{\alpha: \alpha(\eta - \eta') \leq 0} \{ \eta(1 - \eta')\phi(\alpha) + \eta'(1 - \eta)\phi(-\alpha) \}.$$

Motivated from [BJM06]’s work, we define the *classification calibration* of AUC as follows:

Definition 2 *The surrogate loss ϕ is said to be classification-calibrated if*

$$H^-(\eta, \eta') > H(\eta, \eta') \text{ for any } \eta \neq \eta'.$$

Clemencon et al. [CLV08, pp. 846] suggested to study the consistency of AUC through a direct extension of results of Bartlett et al. [BJM06, Theorem 3]; in other words, the following theorem holds for AUC consistency from [BJM06].

Theorem 1 *[BJM06, Theorems 1 and 2] For convex surrogate loss ϕ , the followings are equivalent:*

- ϕ is classification-calibrated.
- ϕ is differential at $t = 0$ and $\phi'(0) < 0$.
- ϕ is consistent with AUC.

This seems that classification calibration completely characterizes the consistency of learning algorithms based on convex surrogate loss, and such results are exactly parallel to those of classification. However, our Lemma 1 shows that this study ignores an important prerequisite: Minimizing the expected ϕ -risk $R_\phi(f)$ over the whole distribution is not equivalent to minimizing the conditional ϕ -risk $C(\mathbf{x}, \mathbf{x}', \alpha)$ on each pair of instances. Therefore, it is not correct to directly use classification calibration to study the consistency of AUC, and as a matter of fact, classification calibration is proven to be a necessary yet insufficient condition for AUC consistency in the next section.

3. AUC Consistency

Recall that

$$R_\phi^* = \inf_f R_\phi(f) = \inf_f E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_X^2} C(\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha),$$

and it is easy to get that

$$R_\phi^* = \inf_f R_\phi(f) \geq E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_X^2} \inf_\alpha C(\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha). \quad (5)$$

It is noteworthy that the equality in Eqn. (5) does not hold for some surrogate losses, which can be shown by the following lemma:

Lemma 1 *For hinge loss $\phi(t) = \max(0, 1 - t)$, it holds that*

$$\inf_f R_\phi(f) > E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_X^2} \inf_\alpha C(\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha).$$

Proof: We prove by contradiction. Suppose that there exists a function f such that

$$R_\phi(f) = E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_X^2} [\inf_\alpha C(\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha)].$$

For simplicity, we consider three different instances $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathcal{X}$ such that

$$\eta(\mathbf{x}_1) < \eta(\mathbf{x}_2) < \eta(\mathbf{x}_3).$$

The conditional risk of hinge loss is given by

$$C(\mathbf{x}, \mathbf{x}', \alpha) = \eta(\mathbf{x})(1 - \eta(\mathbf{x}')) \max(0, 1 - \alpha) + \eta(\mathbf{x}')(1 - \eta(\mathbf{x})) \max(0, 1 + \alpha),$$

and minimizing $C(\mathbf{x}, \mathbf{x}', \alpha)$ gives $\alpha = -1$ if $\eta(\mathbf{x}) < \eta(\mathbf{x}')$. From the assumption that

$$R_\phi(f) = E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_{\mathcal{X}}^2} \inf_{\alpha} C(\eta, \eta', \alpha),$$

we have $f(\mathbf{x}_1) - f(\mathbf{x}_2) = -1$, $f(\mathbf{x}_1) - f(\mathbf{x}_3) = -1$ and $f(\mathbf{x}_2) - f(\mathbf{x}_3) = -1$; while they are contrary to each other. Hence the lemma holds. \square

In a similar manner, we can prove that the following inequality holds for least square loss $\phi(t) = (1 - t)^2$, absolute loss $\phi(t) = |1 - t|$, least square hinge loss $\phi(t) = (\max(0, 1 - t))^2$, etc.,

$$\inf_f R_\phi(f) > E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_{\mathcal{X}}^2} \inf_{\alpha} C(\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha).$$

That is, minimizing the expected ϕ -risk $R_\phi(f)$ over the whole distribution is not equivalent to minimizing the conditional ϕ -risk $C(\mathbf{x}, \mathbf{x}', \alpha)$ on each pair of instances. Therefore, Lemma 1 discloses that, for AUC consistency, we should focus on the expected ϕ -risk over the whole distribution rather than conditional ϕ -risk on each pair of instances. Classification calibration, however, is heavily based on conditional ϕ -risk, and ignores the expected ϕ -risk over the whole distribution; therefore, it is not correct to directly use classification calibration to study AUC consistency.

3.1. Classification Calibration is Necessary yet Insufficient for AUC Consistency

We first prove that hinge loss $\phi(t) = \max(0, 1 - t)$ is inconsistent with respect to AUC by the following theorem:

Theorem 2 *For hinge loss $\phi(t) = \max(0, 1 - t)$, the surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is inconsistent with AUC.*

Proof: For simplicity, we consider three distinct instances $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$, i.e., $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, with marginal probability $\Pr[\mathbf{x}_i] = 1/3$. Further, we set $f_i = f(\mathbf{x}_i)$ and conditional probability $\eta_i = \eta(\mathbf{x}_i)$ such that

$$\eta_1 < \eta_2 < \eta_3, 2\eta_2 < \eta_1 + \eta_3, \text{ and } 2\eta_1 > \eta_2 + \eta_1\eta_3.$$

From Eqn. (3), we have

$$\begin{aligned} R_\phi(f) = & C_0 + C_1\{\eta_1(1 - \eta_2) \max(0, 1 + f_2 - f_1) + \eta_2(1 - \eta_1) \max(0, 1 + f_1 - f_2)\} \\ & + C_1\{\eta_1(1 - \eta_3) \max(0, 1 + f_3 - f_1) + \eta_3(1 - \eta_1) \max(0, 1 + f_1 - f_3)\} \\ & + C_1\{\eta_2(1 - \eta_3) \max(0, 1 + f_3 - f_2) + \eta_3(1 - \eta_2) \max(0, 1 + f_2 - f_3)\}, \end{aligned}$$

where $C_0 = 2(\eta_1 + \eta_2 + \eta_3 - \eta_1^2 - \eta_2^2 - \eta_3^2)/9$ and $C_1 = 2/9$. Minimizing $R_\phi(f)$ gives

$$R_\phi^* = C_0 + C_1(3\eta_1 + 3\eta_2 - 2\eta_1\eta_2 - 2\eta_1\eta_3 - 2\eta_2\eta_3)$$

when $f^* = (f_1^*, f_2^*, f_3^*)$ s.t. $f_1^* = f_2^* = f_3^* - 1$. Notice that the optimal solution should not be $f' = (f'_1, f'_2, f'_3)$ s.t. $f'_1 + 1 = f'_2 = f'_3 - 1$, because

$$\begin{aligned} R_\phi(f') &= C_0 + C_1(2\eta_1(1 - \eta_2) + 3\eta_1(1 - \eta_3) + 2\eta_2(1 - \eta_3)) \\ &= C_0 + C_1(5\eta_1 + 2\eta_2 - 2\eta_1\eta_2 - 3\eta_1\eta_3 - 2\eta_2\eta_3) \\ &= R_\phi^* + C_1(2\eta_1 - \eta_2 - \eta_1\eta_3) > R_\phi^* \end{aligned}$$

where we use the condition $2\eta_1 > \eta_2 + \eta_1\eta_3$.

We now construct a sequence $\{f^{(n)}\}_{n \geq 1}$ by choosing $f^{(1)}(\mathbf{x}_1) = f^{(1)}(\mathbf{x}_2) = f^{(1)}(\mathbf{x}_3) - 1$ and $f^{(n)}(\mathbf{x}) = f^{(1)}(\mathbf{x})$ for $n > 1$. Then, it holds that

$$R_\phi(f^{(n)}) = R_\phi^* \text{ yet } R(f^{(n)}) - R^* = C_1(\eta_2 - \eta_1)/2 \text{ for } n \geq 1.$$

Therefore, there exists a sequence $\{f^{(n)}\}_{n \geq 1}$ such that

$$R_\phi(f^{(n)}) \rightarrow R_\phi^* \text{ yet } R(f^{(n)}) \not\rightarrow R^*,$$

which completes the proof. □

Another relevant loss, the absolute loss $\phi(t) = |1 - t|$, is also proven to be inconsistent with AUC as follows:

Theorem 3 *For absolute loss $\phi(t) = |1 - t|$, the surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is inconsistent with AUC.*

Proof: Similarly to the proof of Theorem 2, we consider $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ with marginal probability $\Pr[\mathbf{x}_i] = 1/3$, and set $f_i = f(\mathbf{x}_i)$ and conditional probability $\eta_i = \eta(\mathbf{x}_i)$ such that

$$\eta_1 < \eta_2 < \eta_3 \text{ and } 2\eta_2 > \eta_1 + \eta_3.$$

From Eqn. (3), we have

$$\begin{aligned} R_\phi(f) = & C_0 + C_1\{\eta_1(1 - \eta_2)|1 + f_2 - f_1| + \eta_2(1 - \eta_1)|1 + f_1 - f_2| + \eta_1(1 - \eta_3)|1 + f_3 - f_1| \\ & + \eta_3(1 - \eta_1)|1 + f_1 - f_3| + \eta_2(1 - \eta_3)|1 + f_3 - f_2| + \eta_3(1 - \eta_2)|1 + f_2 - f_3|\}, \end{aligned}$$

where $C_0 > 0$ and $C_1 > 0$ are independent to f . Minimizing $R_\phi(f)$ gives

$$R_\phi^* = C_0 + C_1(4\eta_1 + \eta_2 + \eta_3 - 2\eta_1\eta_2 - 2\eta_1\eta_3 - 2\eta_2\eta_3)$$

when $f^* = (f_1^*, f_2^*, f_3^*)$ s.t. $f_1^* = f_2^* - 1 = f_3^* - 1$. Notice that the optimal solution should not be $f' = (f_1', f_2', f_3')$ s.t. $f_1' + 1 = f_2' = f_3' - 1$, because

$$\begin{aligned} R_\phi(f') &= C_0 + C_1(2\eta_1(1 - \eta_2) + 3\eta_1(1 - \eta_3) + \eta_3(1 - \eta_1) + 2\eta_2(1 - \eta_3)) \\ &= C_0 + C_1(5\eta_1 + 2\eta_2 + \eta_3 - 2\eta_1\eta_2 - 4\eta_1\eta_3 - 2\eta_2\eta_3) \\ &= R_\phi^* + C_1(\eta_1 + \eta_2 - 2\eta_1\eta_3) > R_\phi^* \end{aligned}$$

where we use $\eta_1 + \eta_2 - 2\eta_1\eta_3 \geq \eta_2 - \eta_1\eta_3 > (\eta_1 + \eta_3)/2 - \eta_1\eta_3 \geq 0$.

We can construct a sequence $\{f^{(n)}\}_{n \geq 1}$ by choosing $f^{(1)}(\mathbf{x}_1) = f^{(1)}(\mathbf{x}_2) - 1 = f^{(1)}(\mathbf{x}_3) - 1$ and $f^{(n)}(\mathbf{x}) = f^{(1)}(\mathbf{x})$ for $n > 1$. Then, it holds that

$$R_\phi(f^{(n)}) = R_\phi^* \text{ yet } R(f^{(n)}) - R^* = C_1(\eta_3 - \eta_2)/2 \text{ for } n \geq 1.$$

Therefore, there exists a sequence $\{f^{(n)}\}_{n \geq 1}$ such that

$$R_\phi(f^{(n)}) \rightarrow R_\phi^* \text{ yet } R(f^{(n)}) \not\rightarrow R^*,$$

which completes the proof. \square

It is noteworthy that hinge loss $\phi(t) = \max(0, 1 - t)$ and absolute loss $\phi(t) = |1 - t|$ are convex and $\phi'(0) = -1 < 0$, and they are classification-calibrated, whereas Theorems 2 and 3 show their inconsistency with AUC, respectively. Therefore, classification calibration is no longer a sufficient condition for AUC consistency.

Corollary 1 *Classification calibration is not sufficient for AUC consistency. For convex function ϕ , the condition that $\phi(t)$ is differential at $t = 0$ with $\phi'(0) < 0$ is not enough for AUC consistency.*

Though classification calibration is not sufficient for AUC consistency, it can be proven to be a necessary condition as follows:

Lemma 2 *If the surrogate loss ϕ is consistent with AUC, then ϕ is classification-calibrated, and for convex ϕ , it is differential at $t = 0$ with $\phi'(0) < 0$.*

The detailed proof is presented in Section 6.1. Based on Corollary 1 and Lemma 2, we derive our first main result:

Theorem 4 *Classification calibration is necessary yet insufficient for AUC consistency.*

This theorem shows that the study on AUC consistency is not similar to that of classification where classification calibration is necessary and sufficient for the consistency of 0/1 loss in [BJM06]. In contrast to Clemencon et al. [CLV08] where hinge loss and absolute loss are consistent with AUC, our results disclose their inconsistency.

3.2. Sufficient Condition for AUC Consistency

In the previous section, we have shown that classification calibration is no longer sufficient for AUC consistency, and therefore, it is necessary to suggest a new sufficient condition. Meanwhile, this new sufficient condition must be based on classification calibration from Lemma 2. We now present a new sufficient condition as follows:

Theorem 5 *The surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is consistent with AUC if $\phi: \mathbb{R} \rightarrow \mathbb{R}$ is a convex, differentiable and non-increasing function with $\phi'(0) < 0$.*

This detailed proof is deferred to Section 6.2. Based on this theorem, it is easy to get:

Corollary 2 *For exponential loss $\phi(t) = e^{-t}$, the surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is consistent with AUC.*

Corollary 3 *For logistic loss $\phi(t) = \ln(1 + e^{-t})$, the surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is consistent with AUC.*

Marron et al. [MTA07] introduced the distance-weighted discrimination method to deal with the problems with high dimension yet small-size sample, and this method has been reformulated by Bartlett et al. [BJM06], for any $\epsilon > 0$, as follows:

$$\phi(t) = \begin{cases} \frac{1}{t} & \text{for } t \geq \epsilon, \\ \frac{1}{\epsilon} \left(2 - \frac{t}{\epsilon}\right) & \text{otherwise.} \end{cases} \quad (6)$$

Based on Theorem 5, we can also derive its consistency as follows:

Corollary 4 *For distance-weighted loss ϕ given by Eqn. (6) with $\epsilon > 0$, the surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is consistent with AUC.*

It is noteworthy that the hinge loss $\phi(t) = \max(0, 1 - t)$ is not differentiable at $t = 1$, and we cannot apply Theorem 5 directly to study the consistency of hinge loss. Theorem 2 proves its inconsistency and also shows the difficulty for consistency without differentiability, even if the surrogate loss function ϕ is convex and non-increasing with $\phi'(0) < 0$. We now derive some variants of hinge loss that are consistent. For example, the q -norm hinge loss:

$$\phi(t) = (\max(0, 1 - t))^q \quad \text{for some } q > 1.$$

Based on Theorem 5, we can get the AUC consistency of the q -norm hinge loss:

Corollary 5 *For q -norm hinge loss $\phi(t) = (\max(0, 1 - t))^q$ with $q > 1$, the surrogate loss $\phi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is consistent with AUC.*

From this corollary, it is immediate to get the consistency for the *least-square hinge loss* $\phi(t) = (\max(0, 1 - t))^2$. We further define the *general hinge loss*, for any $\epsilon > 0$, as:

$$\phi(t) = \begin{cases} 1 - t & \text{for } t \leq 1 - \epsilon, \\ (t - 1 - \epsilon)^2 / 4\epsilon & \text{for } 1 - \epsilon \leq t < 1 + \epsilon, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

It is easy to obtain the AUC consistency of general hinge loss from Theorem 5:

Corollary 6 *For general hinge loss ϕ given by Eqn. (7) with $\epsilon > 0$, the surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is consistent with AUC.*

Hinge loss is inconsistent with AUC, but we can use consistent surrogate loss, e.g., the general hinge loss, to approach hinge loss when $\epsilon \rightarrow 0$. In addition, it is also interesting to derive other surrogate loss functions that are consistent with AUC under the guidance of Theorem 5.

4. Consistent Bounds

4.1. Consistent Bounds for Exponential Loss and Logistic Loss

Corollaries 2 and 3 show that the exponential loss and logistic loss are consistent with AUC, respectively. In this section, we further derive their consistent bounds. The exponential loss and logistic loss possess a special property:

Lemma 3 *For exponential loss and logistic loss, it holds that*

$$\inf_f R_\phi(f) = E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_{\mathcal{X}}^2} \inf_{\alpha} C(\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha).$$

Proof: We provide the detailed proof for the exponential loss, and a similar proof can be obtained for the logistic loss. Fixing an instance $\mathbf{x}_0 \in \mathcal{X}$ and $f(\mathbf{x}_0)$, we set

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \frac{1}{2} \ln \frac{\eta(\mathbf{x})(1 - \eta(\mathbf{x}_0))}{\eta(\mathbf{x}_0)(1 - \eta(\mathbf{x}))} \quad \text{for } \mathbf{x} \neq \mathbf{x}_0.$$

It remains to prove $R(f) = E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_{\mathcal{X}}^2} \inf_{\alpha} C(\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha)$. Based on the above equation, we have, for instances $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$:

$$f(\mathbf{x}_1) - f(\mathbf{x}_2) = \frac{1}{2} \ln \frac{\eta(\mathbf{x}_1)(1 - \eta(\mathbf{x}_2))}{\eta(\mathbf{x}_2)(1 - \eta(\mathbf{x}_1))},$$

which exactly minimizes $C(\eta(\mathbf{x}_1), \eta(\mathbf{x}_2), \alpha)$ when $\alpha = f(\mathbf{x}_1) - f(\mathbf{x}_2)$, and therefore the lemma holds as desired. \square

It is noteworthy that Lemma 3 is constrained to the exponential loss and logistic loss, and it does not hold for other surrogate loss functions such as hinge loss, general hinge loss, q -norm hinge loss, etc. For the exponential loss and logistic loss, Lemma 3 shows that minimizing the expected

risk over the whole distribution is equivalent to minimizing the pairwise-instance conditional risk. Based on this property, we can obtain the consistent bounds for the exponential loss and logistic loss by focusing on their conditional risks. For a general theory, we consider the following equivalence which holds for the exponential loss and logistic loss:

$$\inf_f R_\phi(f) = E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_{\mathcal{X}}^2} \inf_{\alpha} C[\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha],$$

and we denote by f^* the optimal functions, i.e., $R_\phi(f^*) = E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}} \inf_{\alpha} [C(\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha)]$. Under the equivalence assumption, we have

Theorem 6 *For some $c_0 > 0$ and $0 < c_1 \leq 1$, we have*

$$R(f) - R^* \leq c_0 (R_\phi(f) - R_\phi^*)^{c_1},$$

if $(f^(\mathbf{x}) - f^*(\mathbf{x}'))(\eta(\mathbf{x}) - \eta(\mathbf{x}')) > 0$ for $\eta(\mathbf{x}) \neq \eta(\mathbf{x}')$, and*

$$|\eta(\mathbf{x}) - \eta(\mathbf{x}')| \leq c_0 (C(\eta(\mathbf{x}), \eta(\mathbf{x}'), 0) - C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f^*(\mathbf{x}) - f^*(\mathbf{x}')))^{c_1}.$$

This proof is motivated from Zhang [Zha04b] and deferred to Section 6.3. Based on this theorem, we can get the following consistent bounds for the exponential loss and logistic loss:

Corollary 7 *For exponential loss, it holds that $R(f) - R^* \leq \sqrt{R_\phi(f) - R_\phi^*}$.*

Corollary 8 *For logistic loss, it holds that $R(f) - R^* \leq 2\sqrt{R_\phi(f) - R_\phi^*}$.*

The detailed proofs of Corollaries 7 and 8 are given in Section 6.4 and 6.5, respectively.

4.2. Consistent Bounds under Non-Noisy Setting

Now we consider the non-noisy setting [RS09] defined as:

Definition 3 *A distribution \mathcal{D} is said to be non-noisy if it holds either $\eta(\mathbf{x}) = 0$ or $\eta(\mathbf{x}) = 1$ for every $\mathbf{x} \in \mathcal{X}$.*

Under such setting, we have

Theorem 7 For some $c > 0$, we have

$$R(f) - R^* \leq c(R_\phi(f) - R_\phi^*),$$

if $R_\phi^* = 0$, and if $\phi(t) \geq 1/c$ for $t \leq 0$ and $\phi(t) \geq 0$ for $t > 0$.

Proof: For convenience, denote by \mathcal{D}_+ and \mathcal{D}_- the positive and negative instance distributions, respectively. From Eqn. (1), we have

$$R(f) = E_{\mathbf{x} \sim \mathcal{D}_+, \mathbf{x}' \sim \mathcal{D}_-} [I[f(\mathbf{x}) < f(\mathbf{x}')] + I[f(\mathbf{x}) = f(\mathbf{x}')]/2],$$

and thus $R^* = \inf_f [R(f)] = 0$ when $f(\mathbf{x}) > f(\mathbf{x}')$. From Eqn. (3), we get the ϕ -risk $R_\phi(f) = E_{\mathbf{x} \sim \mathcal{D}_+, \mathbf{x}' \sim \mathcal{D}_-} [\phi(f(\mathbf{x}) - f(\mathbf{x}'))]$. Then

$$\begin{aligned} R(f) - R^* &= E_{\mathbf{x} \sim \mathcal{D}_+, \mathbf{x}' \sim \mathcal{D}_-} [I[f(\mathbf{x}) < f(\mathbf{x}')] + I[f(\mathbf{x}) = f(\mathbf{x}')]/2] \\ &\leq E_{\mathbf{x} \sim \mathcal{D}_+, \mathbf{x}' \sim \mathcal{D}_-} [c\phi(f(\mathbf{x}) - f(\mathbf{x}'))] = c(R_\phi(f) - R_\phi^*), \end{aligned}$$

which completes the proof. \square

Based on this theorem, we can get the following corollaries under the non-noisy setting:

Corollary 9 For exponential loss, hinge loss, general hinge loss, q -norm hinge loss, and least square loss $\phi(t) = (1 - t)^2$, we have $R(f) - R^* \leq R_\phi(f) - R_\phi^*$.

Corollary 10 For logistic loss, we have $R(f) - R^* \leq 2(R_\phi(f) - R_\phi^*)$.

It is noteworthy that the hinge loss is consistent with AUC under non-noisy setting although it is inconsistent for the general case as shown in Theorem 2. Moreover, the consistent bounds for the exponential loss and logistic loss under the non-noisy setting are tighter than those of Corollaries 7 and 8, respectively.

5. Equivalence Between Surrogate Losses of AUC and Accuracy

In this section, we study the relationships among AUC, accuracy, and their surrogate loss functions. Our results show that optimizing AUC is more difficult than optimizing accuracy. More

interestingly, we establish an equivalence between the exponential surrogate loss of AUC and the exponential surrogate loss of accuracy regardless of different formulations. This provides a new explanation to the equivalence between AdaBoost and RankBoost: both of them optimize AUC and accuracy simultaneously.

We focus on binary classification and make prediction $y = \text{sgn}[f(\mathbf{x})]$. Thus, optimizing accuracy aims to minimize

$$\begin{aligned} R_{\text{acc}}(f) &= E_{(\mathbf{x}, y) \sim \mathcal{D}} [I[yf(\mathbf{x}) < 0]] \\ &= E_{\mathbf{x}} [\eta(\mathbf{x}) I[f(\mathbf{x}) < 0] + (1 - \eta(\mathbf{x})) I[f(\mathbf{x}) > 0]], \end{aligned}$$

and it is easy to obtain the set of Bayes predictors for accuracy:

$$\mathcal{B}_{\text{acc}} = \{f: f(\mathbf{x})(\eta(\mathbf{x}) - 1/2) > 0 \text{ for } \eta(\mathbf{x}) \neq 1/2\}.$$

Recall that the set of Bayes predictors for AUC from Eqn. (2):

$$\mathcal{B} = \{f: R(f) = R^*\} = \{f: (f(\mathbf{x}) - f(\mathbf{x}'))(\eta(\mathbf{x}) - \eta(\mathbf{x}')) > 0 \text{ if } \eta(\mathbf{x}) \neq \eta(\mathbf{x}')\}.$$

By comparing the two sets of Bayes predictors, we can find that optimizing accuracy tries to learn a function f s.t. $\text{sgn}[f(\mathbf{x})] = \text{sgn}[\eta(\mathbf{x}) - 1/2]$, yet optimizing AUC aims to learn a function which orders instances according to their conditional probability $\eta(\mathbf{x})$. It is easy to construct the Bayes predictor $f_{\text{acc}}^*(\mathbf{x})$ of accuracy from the Bayes predictor $f^*(\mathbf{x})$ of AUC by setting $f_{\text{acc}}^*(\mathbf{x}) = f^*(\mathbf{x}) - f^*(\mathbf{x}_0)$ where $\eta(\mathbf{x}_0) = 1/2$. The converse direction, however, does not hold because we can only order the instances $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ when $\eta(\mathbf{x}) > 1/2 > \eta(\mathbf{x}')$ but fail for $(\eta(\mathbf{x}) - 1/2)(\eta(\mathbf{x}') - 1/2) > 0$. In this sense, it is more difficult to optimize AUC than accuracy.

We consider one of the most popular surrogate loss functions of accuracy:

$$\phi_{\text{acc}}(f(\mathbf{x}), y) = \phi(yf(\mathbf{x}))$$

where ϕ is convex and non-increasing, e.g., the hinge loss $\phi(t) = \max(0, 1 - t)$ [Vap98], exponential loss $\phi(t) = e^{-t}$ [FS97], logistic loss $\phi(t) = \ln(1 + e^{-t})$ [FHT00], etc.

We can also define the ϕ_{acc} -risk as $R_{\phi_{\text{acc}}}(f) = E_{\mathcal{D}}[\phi_{\text{acc}}(f(\mathbf{x}), y)] = E_{\mathcal{D}}[\phi(yf(\mathbf{x}))]$ for accuracy. Since the surrogate loss ϕ_{acc} focuses on single instances, we have

$$\inf_f R_{\phi_{\text{acc}}}(f) = E_{\mathbf{x}} \inf_{f(\mathbf{x})} [C_{\text{acc}}(\eta(\mathbf{x}), f(\mathbf{x}))], \quad (8)$$

where the conditional risk $C_{\text{acc}}(\eta(\mathbf{x}), f(\mathbf{x})) = \eta(\mathbf{x})\phi(f(\mathbf{x})) + (1 - \eta(\mathbf{x}))\phi(-f(\mathbf{x}))$. In other words, minimizing the expected risk over the whole distribution is equivalent to minimizing the conditional risk on every instance. Thus, it is sufficient to study the consistency of accuracy based on conditional risk as done in [BJM06, Zha04b]. This is quite different from our work on AUC consistency. The surrogate loss function for AUC is defined on a pair of instances, and for some surrogate loss functions, minimizing the expected risk over the whole distribution is not be equivalent to minimizing the conditional risk on every pair of instances, as shown by Lemma 1. Therefore, the study on the consistency of AUC is more difficult than the consistency analysis of accuracy.

In what follows, we will study the relationship between the surrogate loss of accuracy, $\phi_{\text{acc}}(f(\mathbf{x}), y) = \phi(yf(\mathbf{x}))$, and the surrogate loss of AUC, $\phi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$, especially for $\phi(t) = e^{-t}$ (exponential loss). The following lemma shows that the exponential surrogate losses of accuracy and AUC have the same optimal solution:

Lemma 4 *The optimal functions of the exponential surrogate loss of accuracy $E_{(\mathbf{x}, y) \sim \mathcal{D}}[e^{-yf(\mathbf{x})}]$ optimize the exponential surrogate loss of AUC*

$$E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}_{\mathbf{X}}^2} [\eta(\mathbf{x})(1 - \eta(\mathbf{x}'))e^{-f(\mathbf{x})+f(\mathbf{x}')} + \eta(\mathbf{x}')(1 - \eta(\mathbf{x}))e^{-f(\mathbf{x}')+f(\mathbf{x})}],$$

and the converse direction holds by fixing $f(\mathbf{x}_0) = 0$ for $\eta(\mathbf{x}_0) = 1/2$.

Proof: From Lemma 3 and Eqn. (8), it suffices to proceed on conditional risk. Minimizing the accuracy's conditional risk $\eta(\mathbf{x})e^{-f(\mathbf{x})} + (1 - \eta(\mathbf{x}))e^{f(\mathbf{x})}$ gives the optimal solution $f_{\text{acc}}^*(\mathbf{x}) = 0.5 \ln(\eta(\mathbf{x})/(1 - \eta(\mathbf{x})))$. On the other hand, minimizing the AUC's conditional risk

$$\eta(\mathbf{x})(1 - \eta(\mathbf{x}'))e^{-f(\mathbf{x})+f(\mathbf{x}')} + \eta(\mathbf{x}')(1 - \eta(\mathbf{x}))e^{-f(\mathbf{x}')+f(\mathbf{x})}$$

gives the optimal solution

$$f^*(\mathbf{x}) - f^*(\mathbf{x}') = 0.5 \ln(\eta(\mathbf{x})(1 - \eta(\mathbf{x}')/\eta(\mathbf{x}')/(1 - \eta(\mathbf{x}))) = f_{\text{acc}}^*(\mathbf{x}) - f_{\text{acc}}^*(\mathbf{x}'),$$

which completes the proof by simple analysis. \square

Similar result also holds for logistic loss $\phi(t) = \ln(1 + e^{-t})$. Based on this lemma, we can further derive the following theorem, whose proof is deferred to Section 6.6.

Theorem 8 *For exponential loss and sequence $\{f^{(n)}\}_{n \geq 1}$, we have $R_\Psi(f^{(n)}) \rightarrow R_\Psi^*$ if $R_{\Psi_{acc}}(f^{(n)}) \rightarrow R_{\Psi_{acc}}^*$; we also have $R_{\Psi_{acc}}(f^{(n)}) \rightarrow R_{\Psi_{acc}}^*$ if $R_\Psi(f^{(n)}) \rightarrow R_\Psi^*$ by setting $f^{(n)}(x_0) = 0$ for $\eta(x_0) = 1/2$ and $n \geq 1$.*

This theorem discloses the asymptotic equivalence between the exponential surrogate loss of accuracy and the exponential surrogate loss of AUC. Thus, the accuracy's surrogate loss $\phi_{acc}(f(\mathbf{x}), y) = e^{-yf(\mathbf{x})}$ is consistent with AUC, whereas the AUC's surrogate loss $\phi(f, \mathbf{x}, \mathbf{x}') = e^{-(f(\mathbf{x})-f(\mathbf{x}'))}$ is consistent with accuracy by choosing a proper threshold. One direct consequence of this theorem is: AdaBoost and RankBoost are equivalent asymptotically, i.e., both of them optimize AUC and accuracy simultaneously for infinite training sample, because AdaBoost and RankBoost essentially optimize the surrogate loss $\phi_{acc}(f(\mathbf{x}), y) = e^{-yf(\mathbf{x})}$ and $\phi(f, \mathbf{x}, \mathbf{x}') = e^{-(f(\mathbf{x})-f(\mathbf{x}'))}$, respectively. It will be interesting to make similar consideration for the logistic loss, and we leave it to future work.

6. Proofs

In this section, we provide some detailed proofs.

6.1. Proof of Lemma 2

Proof: If ϕ is not classification-calibrated, then there exist η_0 and η'_0 such that $\eta_0 > \eta'_0$ (without loss of generality) and $H^-(\eta_0, \eta'_0) = H(\eta_0, \eta'_0)$, i.e.,

$$\begin{aligned} & \inf_{\alpha \in \mathbb{R}} \{ \eta_0(1 - \eta'_0)\phi(\alpha) + \eta'_0(1 - \eta_0)\phi(-\alpha) \} \\ &= \inf_{\alpha: \alpha(\eta_0 - \eta'_0) \leq 0} \{ \eta_0(1 - \eta'_0)\phi(\alpha) + \eta'_0(1 - \eta_0)\phi(-\alpha) \}. \end{aligned}$$

This implies that there is a $\alpha_0 \leq 0$ such that

$$\eta_0(1 - \eta'_0)\phi(\alpha_0) + \eta'_0(1 - \eta_0)\phi(-\alpha_0) = \inf_{\alpha \in \mathbb{R}} \{ \eta_0(1 - \eta'_0)\phi(\alpha) + \eta'_0(1 - \eta_0)\phi(-\alpha) \}.$$

Suppose that the instance space $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2\}$ with marginal probability $\Pr[\mathbf{x}_i] = 1/2$ and conditional probability $\eta(\mathbf{x}_1) = \eta_0$ and $\eta(\mathbf{x}_2) = \eta'_0$. We construct a sequence $\{f^{(n)}\}_{n \neq 1}$ by picking up $f^{(n)}(\mathbf{x}_1) = f^{(n)}(\mathbf{x}_2) + \alpha_0$, and it is easy to get that

$$R_\phi(f^{(n)}) \rightarrow R_\phi^* \text{ yet } R(f^{(n)}) - R^* = (\eta_0 - \eta'_0)/8 \text{ as } n \rightarrow \infty,$$

which implies that ϕ is inconsistent with AUC. Therefore, classification calibration is necessary for AUC consistency.

For classification, Bartlett et al. [BJM06] established that, for convex ϕ , classification calibration is equivalent to the condition that ϕ is differential at $t = 0$ and $\phi'(t) < 0$, whereas no study is provided to guarantee such equivalence for AUC. Therefore, we present the complete proof that, for convex ϕ , the condition that ϕ is differential at $t = 0$ with $\phi'(0) < 0$ is necessary for AUC consistency.

We consider the instance space $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2\}$ with marginal probability $\Pr[\mathbf{x}_1] = \Pr[\mathbf{x}_2] = 1/2$ and conditional probability $\eta(\mathbf{x}_1) = \eta_1$ and $\eta(\mathbf{x}_2) = \eta_2$. We first prove that if the consistent surrogate loss ϕ is differential at $t = 0$, then $\phi'(0) < 0$. Assume $\phi'(0) \geq 0$, and for convex ϕ , we have

$$\begin{aligned} & \eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha) \\ & \geq (\eta_1 - \eta_2)\alpha\phi'(0) + (\eta_1(1 - \eta_2) + \eta_2(1 - \eta_1))\phi(0) \\ & \geq (\eta_1(1 - \eta_2) + \eta_2(1 - \eta_1))\phi(0) \text{ for } (\eta_1 - \eta_2)\alpha \geq 0. \end{aligned}$$

Therefore, we have

$$\begin{aligned} H(\eta_1, \eta_2) &= \inf_{\alpha \in \mathbb{R}} \{ \eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha) \} \\ &= \min \left\{ \inf_{(\eta_1 - \eta_2)\alpha \geq 0} \{ \eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha) \}, \right. \\ &\quad \left. \inf_{(\eta_1 - \eta_2)\alpha \leq 0} \{ \eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha) \} \right\} \\ &= \min \{ \{ \eta_1(1 - \eta_2)\phi(0) + \eta_2(1 - \eta_1)\phi(0) \}, \\ &\quad \inf_{(\eta_1 - \eta_2)\alpha \leq 0} \{ \eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha) \} \} \\ &= \inf_{(\eta_1 - \eta_2)\alpha \leq 0} \{ \eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha) \} \\ &= H^-(\eta_1, \eta_2), \end{aligned} \tag{9}$$

which implies that ϕ is not classification-calibrated. This is contrary to the assumption that ϕ is consistent with AUC from previous analysis,.

We now prove that convex loss ϕ must be differential at $t = 0$ if it is consistent with AUC. Suppose that ϕ is not differential at $t = 0$. Then, we can find subgradients $g_1 > g_2$ such that

$$\phi(t) \geq g_1 t + \phi(0) \text{ and } \phi(t) \geq g_2 t + \phi(0) \text{ for } t \in \mathbb{R},$$

and we consider the following cases:

1. If $g_1 > g_2 \geq 0$, then we choose $\eta_1 = g_1/(g_1 + g_2)$ and $\eta_2 = g_2/(g_1 + g_2)$. It is obvious that $\eta_1 > \eta_2$, and for any $\alpha \geq 0$, we have

$$\begin{aligned} & \eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha) \\ & \geq \eta_1(1 - \eta_2)(g_2\alpha + \phi(0)) + \eta_2(1 - \eta_1)(-g_1\alpha + \phi(0)) \\ & = (\eta_1 g_2 - \eta_2 g_1)\alpha + (g_1 - g_2)\eta_1\eta_2\alpha + (\eta_1(1 - \eta_2) + \eta_2(1 - \eta_1))\phi(0) \\ & = (g_1 - g_2)\eta_1\eta_2\alpha + (\eta_1(1 - \eta_2) + \eta_2(1 - \eta_1))\phi(0) \\ & \geq (\eta_1(1 - \eta_2) + \eta_2(1 - \eta_1))\phi(0); \end{aligned}$$

2. If $g_1 \geq 0 > g_2$ or $g_1 > 0 \geq g_2$, then we choose $\eta_1 = 1$ and $\eta_2 = 1/2$, and for any $\alpha \geq 0$, it holds that

$$\begin{aligned} & \eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha) \\ & \geq \eta_1(1 - \eta_2)(g_1\alpha + \phi(0)) + \eta_2(1 - \eta_1)(-g_2\alpha + \phi(0)) \\ & = g_1\alpha/2 + (\eta_1(1 - \eta_2) + \eta_2(1 - \eta_1))\phi(0) \\ & \geq (\eta_1(1 - \eta_2) + \eta_2(1 - \eta_1))\phi(0); \end{aligned}$$

3. If $0 \geq g_1 > g_2$, then we choose $\eta_1 = (|g_1| + |g_1 - g_2|/2)/(|g_1 + g_2|)$ and $\eta_2 = |g_1|/(|g_1 + g_2|)$.

It is obvious that $\eta_1 > \eta_2$ and for any $\alpha \geq 0$, we have

$$\begin{aligned} & \eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha) \\ & \geq \eta_1(1 - \eta_2)(g_1\alpha + \phi(0)) + \eta_2(1 - \eta_1)(-g_2\alpha + \phi(0)) \\ & = (\eta_1 g_1 - \eta_2 g_2)\alpha + (g_2 - g_1)\eta_1\eta_2\alpha + (\eta_1(1 - \eta_2) + \eta_2(1 - \eta_1))\phi(0) \\ & = (\eta_1(1 - \eta_2) + \eta_2(1 - \eta_1))\phi(0). \end{aligned}$$

Therefore, for any g_1 and g_2 , there exist η_1 and η_2 such that

$$\eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha) \geq (\eta_1(1 - \eta_2) + \eta_2(1 - \eta_1))\phi(0) \text{ for } (\eta_1 - \eta_2)\alpha \geq 0.$$

Similarly to Eqn. (9), we have $H(\eta_1, \eta_2) = H^-(\eta_1, \eta_2)$, and thus ϕ is inconsistent with AUC, which is contrary to the assumption. This completes the proof as desired. \square

6.2. Proof of Theorem 5

We begin with the following lemma, which is crucial to the proof of Theorem 5.

Lemma 5 *For surrogate loss $\phi(f, x, x') = \phi(f(x) - f(x'))$, it holds that*

$$\inf_{f \notin \mathcal{B}} R_\phi(f) > \inf_f R_\phi(f)$$

if $\phi: \mathbb{R} \rightarrow \mathbb{R}$ is a convex, differential and non-increasing function with $\phi'(0) < 0$.

Proof: From the ϕ -risk's definition in Eqn. (3), we have

$$R_\phi(f) = C_0 + \sum_{x, x' \in \mathcal{X}} \Pr[x] \Pr[x'] \left(\eta(x)(1 - \eta(x'))\phi(f(x) - f(x')) + \eta(x')(1 - \eta(x))\phi(f(x') - f(x)) \right)$$

where C_0 is a constant with respect to f . We proceed by contradiction, and suppose that

$$\inf_{f \notin \mathcal{B}} R_\phi(f) = \inf_f R_\phi(f).$$

This implies that there exists an optimal function f^* such that $R_\phi(f^*) = \inf_f R_\phi(f)$ and $f^* \notin \mathcal{B}$, i.e., for some $x_1, x_2 \in \mathcal{X}$, it holds that $f^*(x_1) \leq f^*(x_2)$ yet $\eta(x_1) > \eta(x_2)$.

Since ϕ is convex and differential, the subgradient conditions for minimizing $R_\phi(f)$ give

$$\left[\frac{\partial R_\phi(f)}{\partial f(x_1)} \right]_{f(x_1)=f^*(x_1)} = 0 \quad \text{and} \quad \left[\frac{\partial R_\phi(f)}{\partial f(x_2)} \right]_{f(x_2)=f^*(x_2)} = 0,$$

which are equivalent to

$$\begin{aligned} \sum_{x \neq x_1} \Pr[x] \left(\eta(x_1)(1 - \eta(x))\phi'(f^*(x_1) - f^*(x)) - \eta(x)(1 - \eta(x_1))\phi'(f^*(x) - f^*(x_1)) \right) &= 0 \\ \sum_{x \neq x_2} \Pr[x] \left(\eta(x_2)(1 - \eta(x))\phi'(f^*(x_2) - f^*(x)) - \eta(x)(1 - \eta(x_2))\phi'(f^*(x) - f^*(x_2)) \right) &= 0. \end{aligned}$$

This follows

$$\begin{aligned}
& (\Pr[x_1] + \Pr[x_2]) \left(\eta(x_1)(1 - \eta(x_2))\phi'(f^*(x_1) - f^*(x_2)) - \eta(x_2)(1 - \eta(x_1))\phi'(f^*(x_2) - f^*(x_1)) \right) \\
& + \sum_{x \neq x_1, x_2} \Pr[x] \eta(x) \left((1 - \eta(x_2))\phi'(f^*(x) - f^*(x_2)) - (1 - \eta(x_1))\phi'(f^*(x) - f^*(x_1)) \right) \\
& + \sum_{x \neq x_1, x_2} \Pr[x] (1 - \eta(x)) \left(\eta(x_1)\phi'(f^*(x_1) - f^*(x)) - \eta(x_2)\phi'(f^*(x_2) - f^*(x)) \right) = 0. \quad (10)
\end{aligned}$$

Since ϕ is convex, differential and non-increasing, we have $\phi'(t_1) \leq \phi'(t_2) \leq 0$ when $t_1 \leq t_2$. Therefore, it holds that $\phi'(f^*(x_1) - f^*(x)) \leq \phi'(f^*(x_2) - f^*(x)) \leq 0$ if $f^*(x_1) \leq f^*(x_2)$. This follows

$$\eta(x_1)\phi'(f^*(x_1) - f^*(x)) - \eta(x_2)\phi'(f^*(x_2) - f^*(x)) \leq 0 \quad (11)$$

for $\eta(x_1) > \eta(x_2)$. In a similar manner, we have

$$(1 - \eta(x_2))\phi'(f^*(x) - f^*(x_2)) - (1 - \eta(x_1))\phi'(f^*(x) - f^*(x_1)) \leq 0. \quad (12)$$

For the case $f^*(x_1) = f^*(x_2)$, we have

$$\begin{aligned}
& \eta(x_1)(1 - \eta(x_2))\phi'(f^*(x_1) - f^*(x_2)) - \eta(x_2)(1 - \eta(x_1))\phi'(f^*(x_2) - f^*(x_1)) \\
& = (\eta(x_1) - \eta(x_2))\phi'(0) < 0
\end{aligned}$$

from $\phi'(0) < 0$ and $\eta(x_1) > \eta(x_2)$, which is contrary to Eqn. (10) by combining Eqns. (11) and (12).

For the case $f^*(x_1) < f^*(x_2)$, we have $\phi'(f^*(x_1) - f^*(x_2)) \leq \phi'(0) < 0$ and $\phi'(f^*(x_1) - f^*(x_2)) \leq \phi'(f^*(x_2) - f^*(x_1)) \leq 0$. This follows that, for $\eta(x_1) > \eta(x_2)$,

$$\eta(x_1)(1 - \eta(x_2))\phi'(f^*(x_1) - f^*(x_2)) - \eta(x_2)(1 - \eta(x_1))\phi'(f^*(x_2) - f^*(x_1)) < 0$$

which is also contrary to Eqn. (10) by combining Eqns. (11) and (12). Hence, this lemma follows as desired. \square

Proof of Theorem 5. From Lemma 5, we set

$$\delta = \inf_{f \notin \mathcal{B}} R_\phi(f) - \inf_f R_\phi(f) > 0.$$

Let $\{f^{(n)}\}_{n \geq 0}$ be an any sequence such that $R_\phi(f^{(n)}) \rightarrow R_\phi^*$. Then, there exists an integer $N_0 > 0$ such that

$$R_\phi(f^{(n)}) - R_\phi^* < \delta/2 \text{ for } n \geq N_0.$$

This immediately yields that $f^{(n)} \in \mathcal{B}$ for $n \geq N_0$ from the contrary that

$$R_\phi(f) - R_\phi^* = R_\phi(f) - \inf_{f' \notin \mathcal{B}} R_\phi(f') + \inf_{f' \notin \mathcal{B}} R_\phi(f') - R_\phi^* > \delta \text{ if } f \notin \mathcal{B}.$$

Therefore, we have $R(f^{(n)}) = R^*$ for $n \geq N_0$, which completes the proof. \square

6.3. Proof of Theorem 6

From Eqns. (1) and (2), we have

$$\begin{aligned} & R(f) - R^* \\ &= E_{\eta(x) > \eta(x'), f(x) < f(x')} [\eta(x) - \eta(x')] + E_{\eta(x) > \eta(x'), f(x) = f(x')} [\eta(x)/2 - \eta(x')/2] \\ &\quad + E_{\eta(x) < \eta(x'), f(x) > f(x')} [\eta(x') - \eta(x)] + E_{\eta(x) < \eta(x'), f(x) = f(x')} [\eta(x')/2 - \eta(x)/2] \\ &= E_{(\eta(x) - \eta(x'))(f(x) - f(x')) < 0} [\eta(x) - \eta(x')] + \frac{1}{2} E_{f(x) = f(x')} [\eta(x') - \eta(x)] \\ &\leq E_{(\eta(x) - \eta(x'))(f(x) - f(x')) \leq 0} [\eta(x) - \eta(x')] \\ &\leq E_{(\eta(x) - \eta(x'))(f(x) - f(x')) \leq 0} [c_0 (C(\eta(x), \eta(x'), 0) - C(\eta(x), \eta(x'), f^*(x) - f^*(x')))]^{c_1}, \end{aligned}$$

where the last inequality holds from our assumption. By using the Jensen's inequality, we further obtain

$$R(f) - R^* \leq c_0 \left(E_{(\eta(x) - \eta(x'))(f(x) - f(x')) \leq 0} [C(\eta(x), \eta(x'), 0) - C(\eta(x), \eta(x'), f^*(x) - f^*(x'))] \right)^{c_1}$$

for $0 < c_1 < 1$. This remains to prove that

$$\begin{aligned} & E_{(\eta(x) - \eta(x'))(f(x) - f(x')) \leq 0} [C(\eta(x), \eta(x'), 0) - C(\eta(x), \eta(x'), f^*(x) - f^*(x'))] \\ &\leq E_{(\eta(x) - \eta(x'))(f(x) - f(x')) \leq 0} [C(\eta(x), \eta(x'), f(x) - f(x')) - C(\eta(x), \eta(x'), f^*(x) - f^*(x'))] \\ &= R_\phi(f) - R_\phi^*. \end{aligned}$$

To see it, we consider the following cases:

- If $\eta(x) = \eta(x')$ then $C(\eta(x), \eta(x'), 0) \leq C(\eta(x), \eta(x'), f(x) - f(x'))$ since ϕ is convex;

- If $f(x) = f(x')$ then $C(\eta(x), \eta(x'), 0) = C(\eta(x), \eta(x'), f(x) - f(x'))$;
- If $(\eta(x) - \eta(x'))(f(x) - f(x')) < 0$, then $(f(x) - f(x'))(f^*(x) - f^*(x')) < 0$ from the assumption $(f^*(x) - f^*(x'))(\eta(x) - \eta(x')) > 0$. Thus, 0 is between $f(x) - f(x')$ and $f^*(x) - f^*(x')$, and for convex function ϕ , we have

$$\begin{aligned} C(\eta(x), \eta(x'), 0) &\leq \max(C(\eta(x), \eta(x'), f(x) - f(x')), C(\eta(x), \eta(x'), f^*(x) - f^*(x'))) \\ &= C(\eta(x), \eta(x'), f(x) - f(x')). \end{aligned}$$

Therefore, this theorem follows as desired. \square

6.4. Proof of Corollary 7

For exponential loss $\phi(t) = e^{-t}$, we have the optimal function f^* such that

$$f^*(x) - f^*(x') = \frac{1}{2} \ln \frac{\eta(x)(1 - \eta(x'))}{\eta(x')(1 - \eta(x))} \quad (13)$$

by minimizing the conditional risk $C(\eta(x), \eta(x'), f(x) - f(x'))$, and this follows

$$(f^*(x) - f^*(x'))(\eta(x) - \eta(x')) > 0 \text{ for } \eta(x) \neq \eta(x').$$

From Eqn. (13), we have

$$C(\eta(x), \eta(x'), f^*(x) - f^*(x')) = 2\sqrt{\eta(x)\eta(x')(1 - \eta(x'))(1 - \eta(x))},$$

and it is easy to get $C(\eta(x), \eta(x'), 0) = \eta(x)(1 - \eta(x')) + \eta(x')(1 - \eta(x))$. Therefore, we have

$$\begin{aligned} &C(\eta(x), \eta(x'), 0) - C(\eta(x), \eta(x'), f^*(x) - f^*(x')) \\ &= (\sqrt{\eta(x)(1 - \eta(x'))} - \sqrt{\eta(x')(1 - \eta(x))})^2 \\ &= \frac{|\eta(x) - \eta(x')|^2}{(\sqrt{\eta(x)(1 - \eta(x'))} + \sqrt{\eta(x')(1 - \eta(x))})^2} \\ &\geq |\eta(x) - \eta(x')|^2, \end{aligned}$$

where the last inequality holds from $\eta(x), \eta(x') \in [0, 1]$. Hence, this lemma holds by applying Theorem 6 to exponential loss. \square

6.5. Proof of Corollary 8

For logistic loss $\phi(t) = \ln(1 + e^{-t})$, we have the optimal function f^* such that

$$f^*(x) - f^*(x') = \ln \frac{\eta(x)(1 - \eta(x'))}{\eta(x')(1 - \eta(x))}, \quad (14)$$

by minimizing the conditional risk $C(\eta(x), \eta(x'), f(x) - f(x'))$, and this immediately yields

$$(f^*(x) - f^*(x'))(\eta(x) - \eta(x')) > 0 \text{ for } \eta(x) \neq \eta(x').$$

Therefore, we complete the proof by applying Theorem 6 to logistic loss if the following holds:

$$C(\eta(x), \eta(x'), 0) - C(\eta(x), \eta(x'), f^*(x) - f^*(x')) \geq |\eta(x) - \eta(x')|^2/4. \quad (15)$$

We will prove that Eqn. (15) holds for $|\eta(x') - 0.5| \leq |\eta(x) - 0.5|$, and similar derivation could be made when $|\eta(x') - 0.5| > |\eta(x) - 0.5|$. For notational simplicity, we denote by $\eta = \eta(x)$ and $\eta' = \eta(x')$. Fix η' and we set

$$F(\eta) = C(\eta, \eta', 0) - C(\eta, \eta', f^*(x) - f^*(x')) - (\eta - \eta')^2/4.$$

From Eqn. (14), we further get

$$\begin{aligned} F(\eta) &= \ln(2)(\eta + \eta' - 2\eta'\eta) - (\eta - \eta')^2/4 \\ &\quad - \eta(1 - \eta') \ln \left(1 + \frac{\eta'(1 - \eta)}{\eta(1 - \eta')}\right) - \eta'(1 - \eta) \ln \left(1 + \frac{\eta(1 - \eta')}{\eta'(1 - \eta)}\right). \end{aligned}$$

It is easy to obtain $F(\eta') = 0$ and the derivative

$$\begin{aligned} F'(\eta) &= \ln(2)(1 - 2\eta') - (\eta - \eta')/2 \\ &\quad - (1 - \eta') \ln \left(1 + \frac{\eta'(1 - \eta)}{\eta(1 - \eta')}\right) + \eta' \ln \left(1 + \frac{\eta(1 - \eta')}{\eta'(1 - \eta)}\right). \end{aligned}$$

Further, we have $F'(\eta') = 0$ and the second-order derivative

$$F''(\eta) = \frac{\eta'(1 - \eta')}{\eta(1 - \eta)(\eta + \eta' - 2\eta\eta')} - \frac{1}{2} \geq 0,$$

where the inequality holds since $\eta + \eta' - 2\eta\eta' = \eta(1 - \eta') + \eta'(1 - \eta) < 2$ and $\eta'(1 - \eta') \geq \eta(1 - \eta)$ from assumption $|\eta' - 0.5| \leq |\eta - 0.5|$. Therefore, $F'(\eta)$ is a non-decreasing function, and this yields that

$$F'(\eta) \leq F'(\eta') = 0 \text{ for } \eta \leq \eta', \text{ and } F'(\eta) \geq F'(\eta') = 0 \text{ for } \eta \geq \eta',$$

which implies that $F(\eta) \geq F(\eta') = 0$. Therefore, we complete the proof. \square

6.6. Proof of Theorem 8

We first introduce a lemma for exponential loss as follows:

Lemma 6 *For some $c_0 > 0$, we have*

$$R_\phi(f) - R_\phi^* \leq 4c_0(R_{\phi_{acc}}(f) - R_{\phi_{acc}}^*) \quad (16)$$

if $E_x[(1 - \eta(x))e^{f(x)}] < c_0$; we also have

$$R_{\phi_{acc}}(f) - R_{\phi_{acc}}^* \leq 2\sqrt{R_\phi(f) - R_\phi^*} \quad (17)$$

if $E_x[\eta(x)e^{-f(x)}] = E_x[(1 - \eta(x))e^{f(x)}]$.

Proof: For accuracy's exponential surrogate loss, we have

$$\begin{aligned} R_{\phi_{acc}}(f) - R_{\phi_{acc}}^* &= E_x \left[\eta(x)e^{-f(x)} + (1 - \eta(x))e^{f(x)} - 2\sqrt{\eta(x)(1 - \eta(x))} \right] \\ &= E_x \left[\left(\sqrt{\eta(x)e^{-f(x)}} - \sqrt{(1 - \eta(x))e^{f(x)}} \right)^2 \right], \end{aligned} \quad (18)$$

and similar results holds for AUC's exponential surrogate loss as follows:

$$\begin{aligned} R_\phi(f) - R_\phi^* &= E_{x,x'} \left[\left(\sqrt{\eta(x)(1 - \eta(x'))e^{-f(x)+f(x')}} \right. \right. \\ &\quad \left. \left. - \sqrt{\eta(x')(1 - \eta(x))e^{f(x)-f(x')}} \right)^2 \right]. \end{aligned} \quad (19)$$

For Eqn. (16), we have

$$\begin{aligned} R_\phi(f) - R_\phi^* &\leq 2E_{x'}[(1 - \eta(x'))e^{f(x')}]E_x \left[\left(\sqrt{\eta(x)e^{-f(x)}} - \sqrt{(1 - \eta(x))e^{f(x)}} \right)^2 \right] \\ &\quad + 2E_x[(1 - \eta(x))e^{f(x)}]E_{x'} \left[\left(\sqrt{(1 - \eta(x'))e^{f(x')}} - \sqrt{\eta(x')e^{-f(x')}} \right)^2 \right] \end{aligned}$$

by using the fact

$$\begin{aligned} &\left(\sqrt{\eta(x)(1 - \eta(x'))e^{-f(x)+f(x')}} - \sqrt{\eta(x')(1 - \eta(x))e^{f(x)-f(x')}} \right)^2 \\ &\leq 2(1 - \eta(x'))e^{f(x')} \left(\sqrt{\eta(x)e^{-f(x)}} - \sqrt{(1 - \eta(x))e^{f(x)}} \right)^2 \\ &\quad + 2(1 - \eta(x))e^{f(x)} \left(\sqrt{(1 - \eta(x'))e^{f(x')}} - \sqrt{\eta(x')e^{-f(x')}} \right)^2. \end{aligned}$$

Therefore, Eqn. (16) holds by using $E_x[(1 - \eta(x))e^{f(x)}] \leq c_0$.

From Eqn. (18), we have

$$(R_{\phi_{\text{acc}}}(f) - R_{\phi_{\text{acc}}}^*)^2 \leq E_{x,x'} \left[\left(\sqrt{\eta(x)e^{-f(x)}} - \sqrt{(1-\eta(x))e^{f(x)}} \right)^2 \left(\sqrt{\eta(x')e^{-f(x')}} + \sqrt{(1-\eta(x'))e^{f(x')}} \right)^2 \right].$$

By using $(a+b)^2 \leq 2(a^2+b^2)$, we further get

$$\begin{aligned} & (R_{\phi_{\text{acc}}}(f) - R_{\phi_{\text{acc}}}^*)^2 \\ & \leq 2E_{x,x'} \left[\left(\sqrt{\eta(x)(1-\eta(x'))e^{-f(x)+f(x')}} - \sqrt{\eta(x')(1-\eta(x))e^{f(x)-f(x')}} \right)^2 \right] \\ & \quad + 2E_{x,x'} \left[\left(\sqrt{\eta(x)\eta(x')e^{-f(x)-f(x')}} - \sqrt{(1-\eta(x))(1-\eta(x'))e^{f(x)+f(x')}} \right)^2 \right]. \end{aligned}$$

We complete the proof of Eqn. (17) since the second term in the above is equal to $2(R_{\phi}(f) - R_{\phi}^*)$ from $E_x[\eta(x)e^{-f(x)}] = E_x[(1-\eta(x))e^{f(x)}]$. The lemma follows as desired. \square

Proof of Theorem 8. From Eqn. (18), we have

$$\sqrt{\eta(x)e^{-f^{(n)}(x)}} - \sqrt{(1-\eta(x))e^{f^{(n)}(x)}} \rightarrow 0$$

almost surely as $n \rightarrow \infty$ if $R_{\phi_{\text{acc}}}(f^{(n)}) \rightarrow R_{\phi_{\text{acc}}}^*$. This follows that $E_x[(1-\eta(x))e^{f^{(n)}(x)}] \leq 1$ as $n \rightarrow \infty$, and we complete the first part of Theorem 8 from Eqn. (16).

From Eqn. (19), we have

$$\sqrt{\eta(x)(1-\eta(x'))e^{-f^{(n)}(x)+f^{(n)}(x')}} - \sqrt{\eta(x')(1-\eta(x))e^{f^{(n)}(x)-f^{(n)}(x')}} \rightarrow 0$$

almost surely as $n \rightarrow \infty$ if $R_{\phi}(f^{(n)}) \rightarrow R_{\phi}^*$. This follows that $E_x[\eta(x)e^{-f^{(n)}(x)}] = E_x[(1-\eta(x))e^{f^{(n)}(x)}]$ when $f^{(n)}(x_0) = 0$ for $\eta(x_0) = 0.5$. This completes the second part of Theorem 8 from Eqn. (17). \square

7. Conclusion and Open Problems

AUC (area under ROC curve) is a popular evaluation criterion widely used in diverse learning tasks. Many convex surrogate loss have been explored to optimize AUC owing to its non-convexity and discontinuousness. Therefore, it is important to study the consistency of learning algorithms based on surrogate losses.

Previous study showed that classification calibration is equivalent to AUC consistency, whereas we find that it ignores an important prerequisite: for the pairwise surrogate loss of AUC, minimizing the expected risk over the whole distribution is not equivalent to minimizing the conditional risk on each pair of instances. We disclose that classification calibration is necessary yet insufficient for AUC consistency, e.g., hinge loss and absolute loss are classification-calibrated whereas they are inconsistent with AUC. We provide a new sufficient condition for the asymptotic consistency of learning approaches based on surrogate loss functions, and based on such finding, many surrogate losses are proven to be consistent such as exponential loss, logistic loss, least-square hinge loss, etc. We also derive the consistent bounds for exponential loss and logistic loss, and obtain the consistent bounds for many surrogate loss functions under the non-noise setting. Furthermore, we disclose an equivalence between the exponential surrogate loss of AUC and exponential surrogate loss of accuracy, and one straightforward consequence of such finding is that AdaBoost and RankBoost are equivalent.

Many problems are left to future work. For example, the first open problem is to study the necessity of the condition that ϕ is non-increasing in Theorem 5. It is natural to consider the least square loss $\phi(t) = (1 - t)^2$, which is convex, differential with $\phi'(0) < 0$, yet increasing for $t > 1$. Actually, it is difficult to study the consistency of least square loss, and let us see some simple cases:

- If $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2\}$ with marginal probability $\Pr[\mathbf{x}_i]$ and conditional probability $\eta(\mathbf{x}_i)$ ($i = 1, 2$), then minimizing $R_\phi(f)$ gives the optimal solution $f = (f(\mathbf{x}_1), f(\mathbf{x}_2))$ s.t.

$$f(\mathbf{x}_1) - f(\mathbf{x}_2) = \text{sgn}(\eta(\mathbf{x}_1) - \eta(\mathbf{x}_2)) \text{ for } \eta(\mathbf{x}_1) \neq \eta(\mathbf{x}_2),$$

which implies least square loss is consistent with AUC when $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2\}$.

- If $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ with marginal probability $\Pr[\mathbf{x}_i] = p_i$ and conditional probability $\eta(\mathbf{x}_i) = \eta_i$ ($1 \leq i \leq 3$), then minimizing $R_\phi(f)$ gives the optimal solution $f = (f(\mathbf{x}_1), f(\mathbf{x}_2), f(\mathbf{x}_3)) = (f_1, f_2, f_3)$ such that

$$\begin{aligned} f_1 - f_2 &= (\eta_1 - \eta_2)(p_1(\eta_1 + \eta_3 - 2\eta_1\eta_3) + p_2(\eta_2 + \eta_3 - 2\eta_2\eta_3) + 2p_3(\eta_3 - \eta_3^2))/\Delta \\ f_1 - f_3 &= (\eta_1 - \eta_3)(p_1(\eta_1 + \eta_2 - 2\eta_1\eta_2) + 2p_2(\eta_2 - \eta_2^2) + p_3(\eta_2 + \eta_3 - 2\eta_2\eta_3))/\Delta \\ f_2 - f_3 &= (\eta_2 - \eta_3)(2p_1(\eta_1 - \eta_1^2) + p_2(\eta_1 + \eta_2 - 2\eta_1\eta_2) + p_3(\eta_1 + \eta_3 - 2\eta_1\eta_3))/\Delta, \end{aligned}$$

where $\Delta = p_1(\eta_1 + \eta_2 - 2\eta_1\eta_2)(\eta_1 + \eta_3 - 2\eta_1\eta_3) + p_2(\eta_1 + \eta_2 - 2\eta_1\eta_2)(\eta_2 + \eta_3 - 2\eta_2\eta_3) + p_3(\eta_1 + \eta_3 - 2\eta_1\eta_3)(\eta_2 + \eta_3 - 2\eta_2\eta_3)$. Therefore, least square loss is consistent with AUC when $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$.

- For $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ ($k = 4$ or $k = 5$), we also find the optimal solution f s.t.

$$f(\mathbf{x}_i) - f(\mathbf{x}_j) = (\eta(\mathbf{x}_i) - \eta(\mathbf{x}_j))\Delta_{i,j} \text{ for } i \neq j,$$

where $\Delta_{i,j} > 0$ have very complicated expressions and we omit them here. Therefore, least square loss is also consistent with AUC.

For more general cases, it is reasonable to conjecture that the optimal solution f has the form of $f(\mathbf{x}_i) - f(\mathbf{x}_j) = (\eta(\mathbf{x}_i) - \eta(\mathbf{x}_j))\Delta_{i,j}$, which shows that least square loss is consistent with AUC, whereas we fail to prove and suggest it as a conjecture:

Conjecture 1 *For least square loss $\phi(t) = (1 - t)^2$, the surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is consistent with AUC.*

Another relevant loss function $\phi(t) = |1 - t|^5$ has been suggested by Breiman [Bre99] to design the boosting algorithm *arc-x4*, and it also remains open to study the consistency of surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = |1 - (f(\mathbf{x}) - f(\mathbf{x}'))|^5$ with respect to AUC.

For AUC consistency, we have presented a necessary condition (Lemma 2), i.e., a consistent and convex surrogate loss $\phi(t)$ must be differential at $t = 0$ and $\phi'(0) < 0$; on the other hand, we have also given a sufficient condition (Theorem 5), i.e., surrogate loss ϕ is consistent with AUC if it is differential, convex and non-increasing with $\phi'(0) < 0$. Therefore, an interesting work is to fill the gap between the sufficient condition and necessary condition. It seems difficult to convince the necessity of the condition that ϕ is non-increasing in Theorem 5, and even for least square loss, it still remains open to discuss on its consistency. Therefore, it is a big challenge to find the necessary and sufficient condition for AUC consistency, and we leave it as an open problem.

In addition, our work could motivate the consistency study on other criterions such as recall, precision, F_1 -score, etc.

References

- [Aga12] S. Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *ArXiv:1207.0268*, 2012.
- [AGH⁺05] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the roc curve. *Journal of Machine Learning Research*, 6:393–425, 2005.
- [AM08] N. Ailon and M. Mohri. An efficient reduction of ranking to classification. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 87–98, Helsinki, Finland, 2008.
- [AN09] S. Agarwal and P. Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10:441–474, 2009.
- [AR05] S. Agarwal and D. Roth. Learnability of bipartite ranking functions. In *Proceedings of the 18th Annual Conference on Learning Theory*, pages 16–31, Bertinoro, Italy, 2005.
- [BBB⁺07] M. Balcan, N. Bansal, A. Beygelzimer, D. Coppersmith, J. Langford, and G. Sorkin. Robust reductions from ranking to classification. In *Proceedings of the 20th Annual Conference on Learning Theory*, pages 604–619, San Diego, CA, 2007.
- [BJM06] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.
- [Bre99] L. Breiman. Prediction games and arcing algorithms. *Neural computation*, 11(7):1493–1517, 1999.
- [Bre04] L. Breiman. Some infinity theory for predictor ensembles. *Annals of Statistics*, 32(1):1–11, 2004.
- [BS05] U. Brefeld and T. Scheffer. AUC maximizing support vector learning. In *Proceedings of the 22nd International Conference on Machine Learning Workshop*, Bonn, Germany, 2005.

- [BY03] P. Bühlmann and B. Yu. Boosting with $l - 2$ -loss: Regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.
- [CLV08] S. Cléménçon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *Annals of Statistics*, 36(2):844–874, 2008.
- [CM04] C. Cortes and M. Mohri. AUC optimization vs. error rate minimization. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 313–320. MIT Press, Cambridge, MA, 2004.
- [CMR07] C. Cortes, M. Mohri, and A. Rastogi. Magnitude-preserving ranking algorithms. In *Proceedings of the 24th Annual International Conference on Machine Learning*, pages 169–176, Corvallis, Oregon, 2007.
- [CSS99] W. W. Cohen, R. E. Schapire, and Y. Singer. Learning to order things. *Neural Computation*, 10:243–270, 1999.
- [CV09] S. Cléménçon and N. Vayatis. Overlaying classifiers: a practical approach for optimal ranking. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 313–320. MIT Press, Cambridge, MA, 2009.
- [CVD09] S. Cléménçon, N. Vayatis, and M. Depecker. AUC optimization and the two-sample problem. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 360–368. MIT Press, Cambridge, MA, 2009.
- [CZ08] D. Cossock and T. Zhang. Statistical analysis of Bayes optimal subset ranking. *IEEE Transaction on Information Theory*, 54(11):5140–5154, 2008.
- [DMJ10] J. C. Duchi, L. W. Mackey, and M. I. Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th International Conference on Machine Learning*, pages 327–334, Haifa, Israel, 2010.
- [Ega75] J. Egan. *Signal detection theory and ROC curve*, *Series in Cognition and Perception*. Academic Press, New York, 1975.

- [Elk01] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 973–978, Seattle, WA, 2001.
- [FHOR11] P. A. Flach, J. Hernández-Orallo, and C. F. Ramirez. A coherent interpretation of AUC as a measure of aggregated classification performance. In *Proceedings of the 28th International Conference on Machine Learning*, pages 657–664, Bellevue, WA, 2011.
- [FHT00] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting (with discussions). *Annals of Statistics*, 28(2):337–407, 2000.
- [FISS03] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.
- [FS97] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
- [GZ11] W. Gao and Z.-H. Zhou. On the consistency of multi-label learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 341–358, Budapest, Hungary, 2011.
- [Han09] D. Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77(1):103–123, 2009.
- [HL05] J. Huang and C. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005.
- [HM82] J. Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 1982.
- [HT96] F. Hsieh and B. Turnbull. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics*, 24(1):25–40, 1996.
- [Joa05] T. Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 377–384, 2005.

- [KDH11] W. Kotlowski, K. Dembczynski, and E. Hüllermeier. Bipartite ranking through minimization of univariate loss. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1113–1120, Bellevue, WA, 2011.
- [Lin02] Y. Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):259–275, 2002.
- [MTA07] J. Marron, M. Todd, and J. Ahn. Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271, 2007.
- [PF01] F. J. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.
- [PFK98] F. J. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the 15th International Conference on Machine Learning*, pages 445–453, 1998.
- [RS09] C. Rudin and R. E. Schapire. Margin-based ranking and an equivalence between AdaBoost and RankBoost. *Journal of Machine Learning Research*, 10:2193–2232, 2009.
- [Rud09] C. Rudin. The p -norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10:2233–2271, 2009.
- [Ste05] I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.
- [TB07] A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.
- [UAG05] N. Usunier, M. R. Amini, and P. Gallinari. A data-dependent generalisation error bound for the auc. In *Proceedings of the 22nd International Conference on Machine Learning Workshop on ROC Analysis*, Bonn, Germany, 2005.
- [Vap98] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.
- [XLL09] F. Xia, T. Y. Liu, and H. Li. Top-k consistency of learning to rank methods. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Ad-*

- vances in Neural Information Processing Systems 22*, pages 2098–2106. MIT Press, Cambridge, MA, 2009.
- [XLW⁺08] F. Xia, T. Y. Liu, J. Wang, W. Zhang, and H. Li. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1192–1199, Helsinki, Finland, 2008.
- [Zha04a] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.
- [Zha04b] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85, 2004.
- [ZHJY11] P. Zhao, S. Hoi, R. Jin, and T. Yang. Online AUC maximization. In *Proceedings of the 25th International Conference on Machine Learning*, pages 233–240, Bellevue, WA, 2011.
- [ZOM02] X. Zhou, N. Obuchowski, and D. McClish. *Statistical Methods in Diagnostic Medicine*. John Wiley and Sons, New York, 2002.